Engineering Spatial and Molecular Features from Cellular Niches to Inform Predictions of Inflammatory Bowel Disease

Myles Joshua Toledo Tan[1,2][0000−0002−1426−6526], Maria Kapetanaki[3], and Panayiotis V. Benos[2][0000−0003−3172−3132]

[1] Department of Electrical and Computer Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, Florida, 32611, United States
mylesjoshua.tan@medicine.ufl.edu

[2] Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, Florida, 32610, United States
pbenos@ufl.edu

[3] Department of Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, Florida, 32603, United States of America
mkapetanaki@ufl.edu

Abstract. Differentiating between the two main subtypes of Inflammatory Bowel Disease (IBD): Crohn's disease (CD) and ulcerative colitis (UC) is a persistent clinical challenge due to overlapping presentations. This study introduces a novel computational framework that employs spatial transcriptomics (ST) to create an explainable machine learning model for IBD classification. We analyzed ST data from the colonic mucosa of healthy controls (HC), UC, and CD patients. Using Non-negative Matrix Factorization (NMF), we first identified four recurring cellular niches, representing distinct functional microenvironments within the tissue. From these niches, we systematically engineered 44 features capturing three key aspects of tissue pathology: niche composition, neighborhood enrichment, and niche-gene signals. A multilayer perceptron (MLP) classifier trained on these features achieved an accuracy of 0.774 ± 0.161 for the more challenging three-class problem (HC, UC, and CD) and 0.916 ± 0.118 in the two-class problem of distinguishing IBD from healthy tissue. Crucially, model explainability analysis revealed that disruptions in the spatial organization of niches were the strongest predictors of general inflammation, while the classification between UC and CD relied on specific niche-gene expression signatures. This work provides a robust, proof-of-concept pipeline that transforms descriptive spatial data into an accurate and explainable predictive tool, offering not only a potential new diagnostic paradigm but also deeper insights into the distinct biological mechanisms that drive IBD subtypes.

Keywords: cellular niches · explainable machine learning (XML) · feature engineering · inflammatory bowel disease · spatial transcriptomics.

1 Introduction

Inflammatory Bowel Disease (IBD) presents a significant clinical challenge due to the diagnostic and therapeutic ambiguity between its two main subtypes, Crohn's disease (CD) and ulcerative colitis (UC) [1,2]. Though pathologically distinct, overlapping clinical presentations can complicate diagnosis, which is critical as treatment strategies diverge substantially [3]. Patient heterogeneity in presentation and therapeutic response underscores the need for precise diagnostic tools reflecting underlying cellular and molecular complexity [4].

Single-cell RNA sequencing (scRNA-seq) has provided unprecedented insight into this complexity, revealing diverse immune, stromal, and epithelial cell states in the inflamed gut [5,6,7]. Garrido-Trigo et al. combined scRNA-seq with spatial imaging to highlight that the greatest inter-patient variability in IBD lies within myeloid cells, particularly macrophages and neutrophils [8]. While essential for unraveling disease mechanisms at a cell-by-cell resolution [9], dissociation-based scRNA-seq loses native tissue architecture. Spatial transcriptomics (ST) preserves this, enabling gene expression analysis in its morphological context [10,11]. Here, we use ST to investigate whether the organization of cells into functional microenvironments can serve as a robust diagnostic tool for IBD subtypes.

We hypothesized that disruptions in the structure of *cellular niches*[4], which are localized communities of interacting cells, are distinct hallmarks of UC and CD. Our computational approach classified cell types [12] to map cell populations across ST data from colonic tissue, then applied non-negative matrix factorization (NMF) to identify four recurring cellular niches defined by unique cell type combinations. This analytical framework, which deconstructs spatial data into functional units, has proven effective in identifying pathological microenvironments in other complex inflammatory conditions like idiopathic pulmonary fibrosis [13]. Similar ST approaches have mapped healing programs in mouse models of colitis [14], fibrosis-associated networks in stricturing CD [15], and cellular ecosystems correlating with UC therapeutic response [16]. To build a predictive model, we engineered 44 features capturing three key aspects of the tissue: niche composition (relative abundance), neighborhood enrichment (spatial interactions), and niche-gene signals (localized gene expression). A multilayer perceptron (MLP) classifier was then trained on these features to distinguish between healthy controls (HC), UC, and CD.

Our work builds on the foundational spatial characterizations of IBD [8] and niche decomposition frameworks in fibrosis [13] by introducing a multilayered feature engineering strategy. Although recent landmark studies have focused on characterizing spatial landscapes [14,15] or correlating them with treatment outcomes [16], our approach represents a critical next step: transforming these descriptive spatial insights into a robust predictive model capable of distinguishing IBD subtypes. This approach not only provides a potential new diagnostic

[4] In this article, the term *niche* is used to mean a latent factor. This is a data-derived unit representing recurring patterns of cellular organization inferred from NMF, rather than a predefined biological microenvironment.

paradigm, but also offers deeper insights into the distinct biological mechanisms that drive UC and CD, linking cellular atlases to the context of functional tissue and paving the way for more targeted therapeutic strategies.

Given the methodological complexity of this article, we provide an overview of the computational pipeline in Figure 7 (Appendix).

2 Methods

2.1 Data

The study used two publicly available datasets from the NCBI Gene Expression Omnibus (GEO). The primary dataset, GSE234713 [8,17], consisted of ST data from colonic mucosa. This data was collected from nine formalin-fixed paraffin-embedded (FFPE) human samples, including three healthy non-IBD controls (HC), three patients with CD, and three with UC. Generated using the NanoString CosMx [18] platform, this dataset provided expression profiles for 980 genes. The secondary dataset, GSE214695 [8,19], contained scRNA-seq data from a separate cohort of healthy and IBD colonic mucosa samples. These scRNA-seq data served as a reference for classifying and assigning cell types within the primary spatial dataset. A detailed breakdown of the dataset, including the conditions (health or disease states), FFPE samples per state, number of fields of view (FOV) per sample, and number of cells per sample, is provided in Table 1.

Table 1: Overview of IBD CosMx NanoString data from colonic mucosa

Condition Groups	**FFPE Samples**	**# of FOVs**	**# of Cells**
healthy controls (HC)	HC a	19	39,101
	HC b	20	54,059
	HC c	16	27,905
	HC total	***55***	***121,065***
ulcerative colitis (UC)	UC a	19	49,240
	UC b	22	76,613
	UC c	21	54,811
	UC total	***62***	***180,664***
Crohn's disease (CD)	CD a	19	31,582
	CD b	19	72,440
	CD c	16	53,344
	CD total	***54***	***157,366***
3 Condition Groups	**9 FFPE Samples**	**171 FOVs**	**459,095 Cells**

2.2 Cell type classification

Cell type classification was performed using `cell2location` [12] to infer the probability of individual cell types across the ST dataset. The CosMx dataset and a scRNA-seq reference dataset were first aligned by identifying common genes between the two datasets ($n = 976$ shared genes). Genes with low expression

were removed using the following thresholds: a minimum of five cells expressing the gene, a non-zero mean expression ≥ 1.12 counts, and 3% minimum expression frequency. Mitochondrial genes were excluded to reduce noise.

The `cell2location` [12] regression model was trained on the reference scRNA-seq dataset to estimate a gene expression signature matrix $S_{g\times c}$, where $g = 871$ is the number of genes and $c = 54$ is the number of cell types. Training was performed with 500 epochs, a learning rate of 0.002, and a batch size of 2,500 cells.

The trained model was then applied to the spatial data to solve for $W_{s\times c}$, representing the probability of each cell type c at each spatial location s, by maximizing a Gamma-Poisson likelihood. This produced high-resolution spatial maps of cell type distributions for downstream analysis.

2.3 Cellular niche decomposition

Cellular niches were identified by applying NMF [20] to the inferred cell type probability matrix $W_{s\times c}$. The goal was to decompose this matrix into two low-rank, non-negative matrices:

$$W_{s\times c} \approx U_{s\times k} \cdot H_{k\times c}$$

where k is the number of latent factors (niches), U represents the contribution of each spatial location to a niche, and H represents the relative composition of cell types within each niche.

The number of factors was set to $k = 4$, based on the elbow method applied to the reconstruction error curve. The factorization was optimized using the non-negative double singular value decomposition (NNDSVD) [21] initialization with a maximum of 1,000 iterations and a random seed of 0 for reproducibility.

Each cell was assigned to its dominant niche by $\arg\max(U_{s,k})$. This process reveals the cellular composition of each of the four niches (latent factors) and allows for the assignment of each individual cell to its dominant niche, providing a basis for subsequent feature engineering.

2.4 Feature engineering

A set of 44 features that was used for downstream classification consisted of three groups: (i) four niche composition features representing the relative abundance of each niche within a FOV, (ii) 16 neighborhood enrichment features capturing the spatial relationships between niches, and (iii) 24 niche–gene features identified through an information-theoretic selection process. Together, these features comprehensively represent the biological, spatial, and molecular characteristics of each FOV and served as input to the MLP classifier described in a later section.

Niche composition. Each cell was assigned to one of four NMF-derived niches. For each field-of-view (FOV), the niche composition features were computed as the proportion of cells in each niche:

$$\text{comp_niche}_i = \frac{n_i}{\sum_{k=1}^{4} n_k},$$

where n_i is the number of cells in Niche $i \in \{1, 2, 3, 4\}$. This produced four features per FOV.

Niche neighborhood enrichment score To capture local spatial organization, we examined cell–cell neighborhoods within each FOV. A KD-tree [22] was used to identify neighbors, where any cell within a distance equal to twice the diameter of the focal cell was considered a neighbor (Figure 1). Let $O_{i,j}$ be the observed count of ordered neighbor pairs in which the focal cell belongs to Niche i and the neighboring cell to Niche j ($i, j \in \{1, 2, 3, 4\}$). Let p_i represent the overall proportion of Niche i cells in the FOV and let $T = \sum_{i=1}^{4} \sum_{j=1}^{4} O_{i,j}$ be the total number of ordered pairs. The expected count under random mixing is $E_{i,j} = T \cdot p_i p_j$. A Laplace-smoothed log-ratio was then calculated to quantify enrichment $\mathrm{S}_{i,j}$:

$$\mathrm{S}_{i,j} = \log_2 \left(\frac{O_{i,j} + 1}{E_{i,j} + 1} \right).$$

This generated a total of 16 features per FOV (a 4×4 matrix including self-pairs).

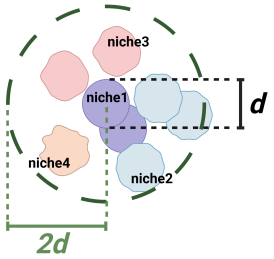


Fig. 1: Illustration of cell–cell neighborhood centered on a focal cell (Niche 1, purple). The focal cell has diameter d, so any cell within a radius $2d$ from the center of the focal cell is part of its neighborhood. In this example, the neighborhood includes one additional Niche 1 cell, three Niche 2 cells (blue), two Niche 3 cells (pink), and one Niche 4 cell (orange). The dashed green circle represents the neighborhood boundary. Created in https://BioRender.com.

Niche–gene features with information-theoretic selection For each niche, gene-level signals were aggregated within every FOV by taking the mean expression of each gene across all cells assigned to that niche, producing features labeled as `niche_[niche#]_gene_[gene]`. To identify the most informative subset of these features for distinguishing among the three condition groups (HC,

UC, and CD) we used mutual information (MI) [23] between each feature X and a binary class label Y_d (1 for the condition group d, 0 otherwise).

For discrete variables, the MI is:

$$I(X;Y_d) = \sum_{x} \sum_{y \in \{0,1\}} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\,p(y)} \right),$$

where $p(x,y)$ is the joint probability of feature value x and label y, while $p(x)$ and $p(y)$ are the corresponding marginal probabilities.

The top 15 features with the highest MI were selected for each group (45 total). For each selected feature, a two-sided Mann–Whitney U test [24,25] was then used to compare its distribution between FOVs labeled as d and those labeled as non-d. Multiple testing correction was performed using the Benjamini–Hochberg false discovery rate (FDR) [26], and any features appearing in more than one disease list were removed to ensure disease specificity. This process yielded 24 unique, statistically significant niche–gene features.

2.5 Multilayer perceptron

A MLP [27,28] was trained to classify each field of view (FOV) into one of three condition groups: HC, UC, or CD. The model input consisted of 44 engineered features ($p = 44$), and the output was a three-class categorical variable ($k = 3$). A pipeline was constructed to standardize features followed by MLP classification. Hyperparameter optimization [29] was performed using randomized search over 30,000 candidate configurations with a three-fold stratified group cross-validation scheme ($n = 171$ FOVs) with patient ID as the grouping variable, i.e., all FOVs from a given patient were assigned to the same fold. The search explored activation functions (ReLU [30] or tanh [31]), regularization strengths $\alpha \sim \text{LogUniform}(10^{-5}, 10^{-1})$, batch sizes $\{2, 4, 8, 16\}$, and seven hidden layer architectures: (25) [single small layer], $(32, 16, 8)$ [progressively tapering], $(40, 20, 10, 5)$ [deeper with small neurons], $(44, 22)$ [starting at feature count, then tapering], $(50, 25, 12)$ [moderately wide tapering], $(50, 50)$ [uniform width], and $(64, 32)$ [wide, shallow network].

The optimal architecture (Figure 2) consisted of four hidden layers with the number of neurons decreasing in size: $40 \to 20 \to 10 \to 5$. It used the ReLU activation function, a batch size of 4, and an L^2 regularization [32] parameter $\alpha = 0.001$ (rounded; results not sensitive to finer precision). This configuration achieved a mean F1-score of 0.712 for the three-class problem. The model was optimized with Adam [33] using `MLPClassifier` in scikit-learn with `learning_rate='adaptive'`, which reduces the base learning rate when training plateaus. This is the scheduler implemented by `MLPClassifier` and is distinct from the per-parameter adaptation inherent to Adam [33]. A maximum of 1000 iterations was used. Performance was evaluated using weighted F1-score as the primary metric, with additional reporting of accuracy, precision, and recall [34]. Final evaluation was based on mean performance across folds and included a confusion matrix to assess misclassification patterns across condition groups.

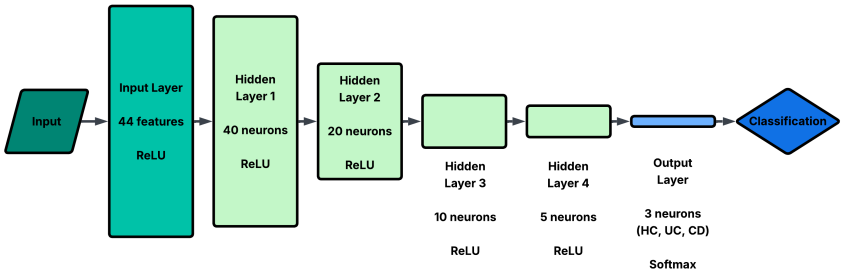


Fig. 2: Architecture of the multilayer perceptron (MLP) used for classification. The network consists of an input layer with 44 features, four hidden layers with 40, 20, 10, and 5 neurons respectively, and an output layer with three neurons (corresponding to HC, UC, and CD) using the softmax activation function.

2.6 Model explainability analysis

Causal discovery Potential cause-effect relationships among features and condition group (Disease/Health State) were inferred using the Fast Causal Inference (FCI)-Stable algorithm [35,36] implemented in the `rCausalMGM` R package [37,38]. Three separate datasets were analyzed: (i) niche composition (four variables representing the log-transformed proportions of each niche per field of view), (ii) 16 neighborhood enrichment features, and (iii) 24 niche–gene features identified through the information-theoretic approach. Each dataset included the categorical condition group variable HC, UC, CD) as a target node.

For each dataset, FCI-Stable was run with a significance level of $\alpha = 0.05$ and the orientation rule set to `maxp` [39]. This algorithm identifies a partial ancestral graph (PAG), where nodes represent either spatial niches, enrichment variables, or niche–gene features, and edges encode potential direct or conditional dependencies. Analyses were performed separately for each condition group and for the combined disease variable, yielding graphical models that capture directional relationships between features and disease. The resulting causal graphs were exported as SIF files for downstream visualization and interpretation.

Statistical tests To assess whether the composition of niches differed across the three groups, the percentage of cells belonging to each NMF factor was calculated for every field of view (FOV). Because the data did not follow a normal distribution, non-parametric statistical methods were employed. Pairwise group comparisons were performed using Dunn's post-hoc test to evaluate differences between HC, UC, and CD for each niche factor. The Benjamini–Hochberg FDR [26] correction was applied to account for multiple testing. For each comparison, the difference in group means was computed, and the direction of change was indicated as "Up" if Group1 > Group2 or "Down" otherwise. Comparisons with adjusted p-values < 0.05 were considered statistically significant.

Neighborhood enrichment was quantified for each FOV by comparing observed versus expected counts of adjacent niche interactions. These enrichment

scores were then analyzed to determine whether the spatial organization of niches differed across HC, UC, and CD. A global Kruskal–Wallis test [40] was first performed for each niche interaction to evaluate overall differences among the three condition groups. When a global test reached significance ($p < 0.05$), pairwise Mann–Whitney U tests [24,25] were subsequently conducted to identify specific group-level differences. The Benjamini–Hochberg procedure [26] was again used to correct for multiple comparisons. Significant pairwise comparisons were reported with their adjusted p-values, and the relative direction of change ("Up" or "Down") indicated whether the enrichment of a given niche interaction was higher or lower in one group compared to another.

Feature importance analysis To interpret the MLP classifier, permutation importance (PI) [41,42] was computed to quantify the contribution of each feature to model performance. For each feature, its values were randomly permuted across samples while keeping other features constant. The resulting decrease in weighted F1-score was recorded. This procedure was repeated multiple times to estimate the mean and standard deviation (SD) of the F1-score change.

The absolute value of the mean decrease was used to rank features, with the sign of the original mean indicating whether the feature had a positive (blue) or negative (red) association with correct classification. Error bars represented the variability (SD) across permutations.

This analysis was performed for both the three-class task (HC, UC, CD) and a two-class task where UC and CD were combined into a single IBD class. The three-class results identified features distinguishing all three conditions, while the binary classification highlighted features most critical for separating HC from IBD. Together, these analyses provided complementary insights into the most influential biological and spatial predictors among the 44 input features.

3 Results

3.1 Cell type classification and cellular niche decomposition

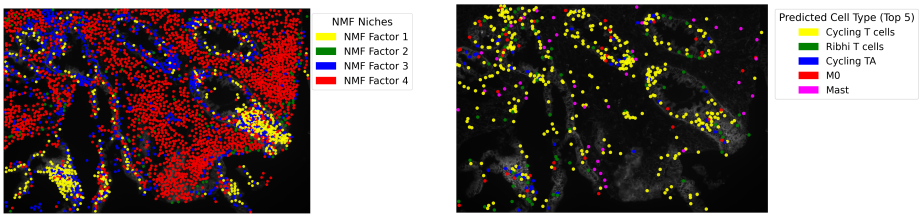
Figure 3 illustrates how our pipeline links computational niche discovery with biological interpretation using one representative field of view (FOV), `UC a_8`.

In panel (a), each point represents a single cell overlaid on the raw histomorphology image. Colors indicate NMF factors, which define distinct cellular niches. This visualization shows how niches are spatially arranged within the tissue and how they relate to the underlying morphology. The color legend allows straightforward identification of niche boundaries and neighboring regions.

Panel (b) focuses on Niche 3 identified in panel (a). Here, cells are colored by their predicted cell type, with only the five most abundant cell types shown. Each color represents a unique cell type, as indicated in the legend, revealing how these populations are distributed within the niche.

Together, these panels demonstrate how computational results can be anchored in biological context. Panel (a) shows the spatial layout of inferred niches,

while panel (b) links one niche to its cellular composition. Although only one FOV is presented, this approach can be applied across larger datasets to uncover biologically meaningful spatial patterns and guide downstream interpretation.



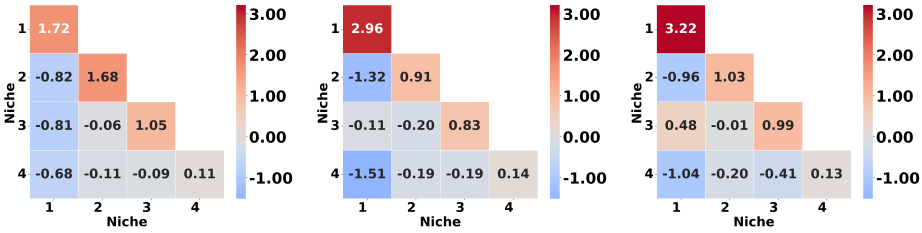
(a) (b)

Fig. 3: Visualization of cellular niches and their cell type composition in FOV `UC a_8`. (a) Cellular niches identified by NMF as colored points overlaid on the histomorphology; (b) The five most abundant cell types within Niche 3.

3.2 Feature engineering

Niche neighborhood enrichment score Figure 4 shows niche neighborhood enrichment scores across the three condition groups. Each heatmap compares observed versus expected interactions between pairs of niches within a given FOV. Red indicates niche pairs that co-occur more frequently than expected, while blue indicates niche pairs that are less frequently observed. The diagonal elements reflect interactions within the same niche, while off-diagonal elements capture relationships between different niches.

This analysis highlights how the spatial organization of cellular niches varies across condition groups. For example, Niche Pair `1,3` is enriched in CD but not in HC and healthy tissue, suggesting disease-specific alterations in tissue structure. These enrichment patterns provide insight into how disease progression reshapes the spatial context of the cellular microenvironment.



(a) (b) (c)

Fig. 4: Niche enrichment comparisons across (a) HC, (b) UC, (c) CD.

Niche–gene features with information-theoretic selection A total of 24 niche-gene features were identified (Table 2). HC-associated features were re-

lated to the genes *PIGR*, *IGHG2*, *HLA-DRB1*, *IGHG1*, *CD38*, *STAT1*, and *CHI3L1* expressed in Niche 1, 2, and 4 cells; while UC-associated features were related to the genes *IGKC*, *IGFBP5*, *HLA-DQB1*, *CASP3*, and *COL3A1* expressed in Niche 2, 3, and 4 cells; and CD-associated features were related to the genes *MZT2A*, *COL9A2*, *FZD1*, *HDAC5*, *COL5A1*, *SOX6*, *FOXF1* expressed in Niche 3 and 4 cells. These features were integrated with niche composition and neighborhood enrichment features to form the final 44-feature dataset used for classification. Table 2 also indicates whether each gene has been previously reported in the literature to be associated with IBD. When marked "yes," the corresponding reference cites the specific study supporting that association.

Table 2: Significant niche-gene features selected through MI score analysis

Niche-Gene Feature	**MI Score**	**Condition Group**	**Adj. *p*-value**	**Known IBD Association**
niche_1_gene_HLA-DRB1	0.351	HC	0.037	yes [43]
niche_1_gene_IGHG2	0.351	HC	<0.001	yes [44]
niche_1_gene_PIGR	0.357	HC	<0.001	yes [45]
niche_2_gene_IGFBP5	0.271	UC	<0.001	yes [46]
niche_2_gene_IGHG1	0.348	HC	<0.001	yes [44]
niche_2_gene_IGHG2	0.352	HC	<0.001	yes [44]
niche_3_gene_COL9A2	0.230	CD	0.020	no
niche_3_gene_HLA-DQB1	0.261	UC	<0.001	yes [47]
niche_3_gene_IGFBP5	0.276	UC	<0.001	yes [46]
niche_3_gene_MZT2A	0.244	CD	<0.001	no
niche_4_gene_CASP3	0.280	UC	<0.001	yes [48]
niche_4_gene_CD38	0.401	HC	<0.001	yes [49]
niche_4_gene_CHI3L1	0.363	HC	<0.001	yes [50]
niche_4_gene_COL3A1	0.267	UC	<0.001	yes [51]
niche_4_gene_COL5A1	0.252	CD	<0.001	yes [52]
niche_4_gene_FOXF1	0.211	CD	0.011	no
niche_4_gene_FZD1	0.352	CD	<0.001	yes [53]
niche_4_gene_HDAC5	0.262	CD	0.004	yes [54]
niche_4_gene_HLA-DQB1	0.277	UC	<0.001	yes [47]
niche_4_gene_IGHG2	0.369	HC	<0.001	yes [44]
niche_4_gene_IGKC	0.313	UC	0.001	yes [55]
niche_4_gene_MZT2A	0.209	CD	0.002	no
niche_4_gene_SOX6	0.235	CD	0.010	yes [56]
niche_4_gene_STAT1	0.367	HC	<0.001	yes [57]

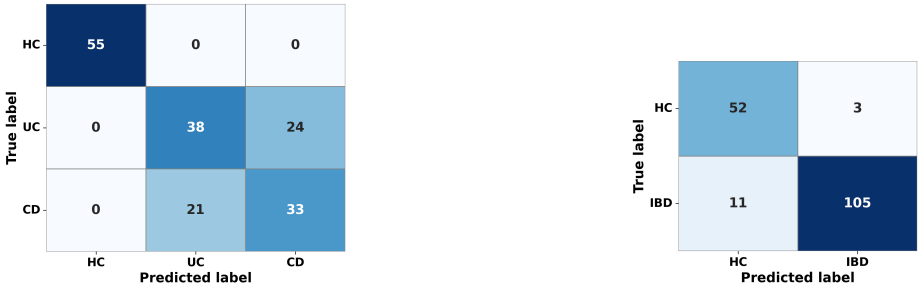
3.3 Multilayer perceptron

Figure 5 shows the confusion matrices for both classification tasks, obtained by aggregating predictions across all three folds of group-stratified cross-validation (patients were held out by group in each fold). In the three-class case (Figure 5a), HC samples were classified perfectly, while misclassifications primarily occurred between UC and CD, indicating overlapping features between these two disease

states. In the two-class case (Figure 5b), the model showed strong separation between HC and IBD, with only a small number of misclassifications.

The corresponding classification reports, computed from the aggregated cross-validation results, are provided in Tables 3 and 4. Table 5 summarizes the mean and standard deviation of each performance metric across the three cross-validation folds, providing a measure of variability in model performance. For the more challenging three-class problem, the macro-averaged F1-score was 0.741 (accuracy 0.774 ± 0.161), with perfect HC classification and errors between UC and CD, reflecting their biological similarity. In comparison, the two-class problem showed excellent separation of HC and IBD, with overall accuracy of 0.916 ± 0.118.

Overall, these results indicate that while distinguishing between UC and CD remains challenging due to biological similarity, the MLP classifier is highly effective at differentiating IBD patients from HC.



(a) (b)

Fig. 5: Confusion matrices for the (a) three-class, and (b) two-class problems.

Table 3: Classification report for the three-class problem (HC, UC, CD).

Class	**Precision**	**Recall**	**F1-Score**	**Support**
HC	1.000	1.000	1.000	55
UC	0.644	0.613	0.628	62
CD	0.579	0.611	0.595	54
Accuracy			0.737	171
Macro Avg	0.741	0.741	0.741	171
Weighted Avg	0.738	0.737	0.737	171

Table 4: Classification report for the two-class problem (HC, IBD)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
HC	0.825	0.945	0.881	55
IBD	0.972	0.905	0.938	116
Accuracy			0.918	171
Macro Avg	0.899	0.925	0.909	171
Weighted Avg	0.925	0.918	0.919	171

Table 5: Three-fold stratified group cross-validation performance for three-class (HC, UC, CD) and two-class (HC vs. IBD) problems.

Performance Metric (Mean ± Std. Dev.)	**Three classes (HC, UC, CD)**	**Two classes (HC, IBD)**
Accuracy	0.774 ± 0.161	0.916 ± 0.118
Precision	0.743 ± 0.220	0.927 ± 0.104
Recall	0.741 ± 0.183	0.908 ± 0.129
F1 Score	0.712 ± 0.209	0.912 ± 0.125

3.4 Model explainability analysis

Causal discovery The causal graph in Figure 6a shows that condition (Disease/Health State) is linked to Niche 1 and Niche 2 compositions, with edges suggesting either influence of disease on these niches or shared unmeasured causes. Niches 3 and 4 appear upstream of Niche 1, indicating that changes in these niches may precede or regulate downstream alterations. The relationship between Niches 3 and 4 remains unresolved, suggesting potential feedback or confounding.

In Figure 6b, two spatial interaction features have direct connections to disease status, highlighting specific niche–niche relationships as strong indicators of health versus disease. Other edges between enrichment features suggest a structured network of spatial dependencies, with some interactions influencing others and a few connections possibly driven by latent variables.

Finally, in Figure 6c, multiple niche–gene features show direct or confounded links to disease, as well. Disease status has bidirected connections with `Niche4_CD38` and `Niche4_STAT1`, as well as partially oriented edges with `Niche3_HLA.DQB1`, `Niche4_COL3A1`, and `Niche1_PIGR`, indicating closely linked associations. Disease status also appears to be directly caused by `Niche4_CASP3`. Moreover, within Niche 4, several features form a densely connected subnetwork. The overall structure indicates a complex network of interrelated niche–gene features involving multiple niches, genes and the condition (Disease/Health State).

Statistical tests Table 6 shows a significantly higher proportion of Niche 1 cells in HC than in UC ($p \approx 0.0003$) and CD ($p < 0.0001$); a significantly higher proportion of Niche 2 cells in UC than in HC ($p < 0.0001$) and CD ($p \approx 0.0219$); a significantly lower proportion of Niche 2 cells in HC than in UC ($p < 0.0001$) and CD ($p < 0.0001$); a significantly higher proportion of Niche 3 cells in UC than in HC ($p < 0.0001$) and CD ($p < 0.0001$); and a significantly higher proportion of Niche 4 cells in CD than in HC ($p < 0.0001$) and UC ($p < 0.0001$).

Table 7 shows a significantly lower $S_{1,1}$ in UC than in HC ($p = 0.032$) and CD ($p = 0.012$); a significantly higher $S_{1,2}$ in HC than in UC ($p < 0.001$) and CD ($p = 0.014$); and a significantly higher $S_{2,1}$ in HC than in UC ($p = 0.001$) and CD ($p = 0.014$).

Only statistically significant results are presented in both tables, with non-significant findings omitted for clarity.

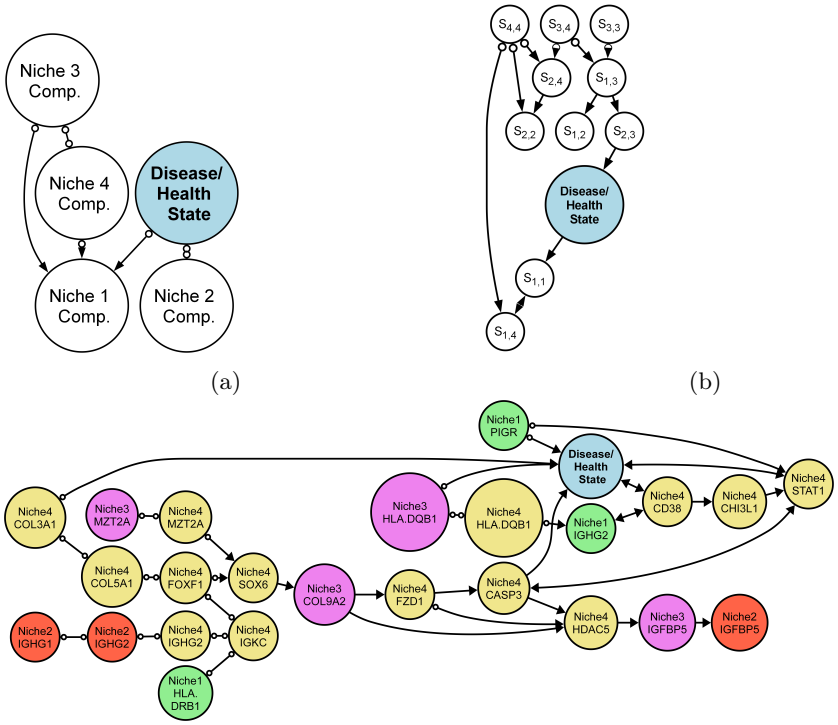


Fig. 6: Causal graphs depicting relationships between features and condition (Disease/Health State): (a) Niche composition, (b) Enrichment score, and (c) Niche-gene features. An open circle in an edge represents the potential presence of a latent confounder.

Table 6: Pairwise comparisons of niche compositions across condition groups.

NMF Factor	Group 1	Group 2	Difference Between Means	Direction*	*p*-value	Observation
1	HC CD	UC HC	7.2503 -10.2075	Up Down	0.0003 <0.0001	HC-associated increase
2	HC CD CD	UC HC UC	-4.4167 2.4362 -1.9805	Down Up Down	<0.0001 <0.0001 0.0219	HC-associated decrease; UC-associated increase
3	HC CD	UC UC	-5.9405 -7.3289	Down Down	<0.0001 <0.0001	UC-associated increase
4	CD CD	HC UC	9.1598 12.2667	Up Up	<0.0001 <0.0001	CD-associated increase

***Note:** "Up" means Group 1 > Group 2.

Feature importance analysis Among the 44 input features, only the top 20 are shown in Figure 8 (Appendix). In the three-class case, niche–gene features dominated, led by `niche_3_gene_MZT2A`, followed by several Niche 4 genes (`COL5A1`, `MZT2A`, `FZD1`, `CASP3`). Other high-ranking features included `niche_-`

Table 7: Pairwise comparisons of niche interactions across condition groups.

Interaction	Group 1	Group 2	Direction*	Group 1 Mean	Group 2 Mean	Adj. *p*-value	Observation
1 vs 1	HC UC	UC CD	Up Down	0.165 -2.627	-2.627 -0.039	0.032 0.012	UC-associated decrease
1 vs 2	HC HC	UC CD	Up Up	-1.308 -1.308	-3.652 -2.650	<0.001 0.014	HC-associated increase
1 vs 4	HC	UC	Up	-1.612	-4.851	0.002	None
2 vs 1	HC HC	UC CD	Up Up	-1.693 -1.693	-3.673 -2.917	0.001 0.014	HC-associated increase
4 vs 1	HC	UC	Up	-1.736	-4.823	0.002	None

***Note:** "Up" means Group 1 > Group 2.

`3_gene_HLA-DQB1`, `niche_1_gene_PIGR`, and `niche_4_gene_COL3A1`, with few enrichment or composition variables (`comp_niche_1`, `comp_niche_4`) appearing. For the two-class task (HC vs. IBD), enrichment features dominated, especially `enrichment_2-2` ($S_{2,2}$), `enrichment_3-1` ($S_{3,1}$), and `enrichment_2-3` ($S_{2,3}$), while only a few niche–gene features ranked highly. Thus, separating UC from CD relied more on gene-level signals, whereas distinguishing HC from IBD depended on spatial interaction patterns.

However, PI can be misleading under correlated or redundant features, especially for highly nonlinear learners (e.g., MLPs). In such cases, permuting a single correlated feature may over- or under-estimate its contribution due to shared information and interaction terms. We therefore interpret PI qualitatively and report it with its variability across permutations, noting this caveat.

4 Discussion

We sought to develop a MLP classifier capable of classifying IBD subtypes by engineering a multi-layered feature set from ST data. The model achieved $\sim 77\%$ accuracy in the more challenging three-class problem, with perfect classification of HC and lower accuracy between UC and CD. This is expected since distinguishing between UC and CD is challenging. A central finding of our work, revealed by PI analysis, is the differential utility of our engineered features. PI analysis revealed that this task depended primarily on niche-gene features, highlighting subtype-specific molecular signals within these microenvironments. In comparison, the simpler two-class problem achieved $\sim 92\%$ accuracy, which was derived mainly from niche neighborhood enrichment scores, indicating that spatial cellular organization is a strong marker of general intestinal inflammation.

Our computational framework moves beyond a *black box* prediction. Its explainable nature, incorporating causal discovery and feature importance analyses, allows for direct biological interpretation of the features driving IBD classification. The reliance of our model on neighborhood enrichment to identify general inflammation provides quantitative evidence that the disruption of spatial cytoarchitecture is a fundamental hallmark of the disease. This work builds upon previous studies that have provided descriptive characterizations of the IBD spatial landscape by transforming these insights into a robust predictive model. The identification of distinct niche-gene features, such as *CASP3* in Niche 4 cells and *HLA-DQB1* in Niches 3 and 4 cells for UC, and collagen-associated genes

like *COL5A1* in Niche 4 cells for CD, offers new testable hypotheses about the specific molecular pathways that define these subtypes.

The primary strength of our methodology lies in its novel, three-tiered feature engineering strategy, which captures niche composition, spatial enrichment, and localized gene expression to provide a holistic view of the tissue state. The use of NMF to define cellular niches provided an unbiased, data-driven approach to deconstructing complex spatial data into functional units. However, the small cohort of nine samples (three per condition) limits generalizability. The analysis was also constrained by the 980-gene panel; a whole-transcriptome approach could reveal more biomarkers. In addition, we recognize the risk of overfitting from extensive hyperparameter tuning on small datasets [58]. To mitigate this, we employed patient-level group-stratified cross-validation, ensuring that FOVs from the same patient never appeared in both training and validation folds. We also used L^2 regularization and a compact feature space to constrain complexity. Nonetheless, given the limited number of patients, the reported accuracies should be interpreted as cross-validated estimates within a proof-of-concept framework, pending external validation on larger, independent, and multi-center cohorts.

This framework has significant clinical and research implications. It offers a potential path toward a more objective, data-driven diagnostic tool to help resolve the ambiguity between UC and CD. The specific spatial patterns and molecular markers identified could form the basis of novel diagnostic assays. For research, our findings provide deeper insights into the distinct biological mechanisms driving these diseases. The causal graphs suggest testable hypotheses about how changes in niche composition and interactions drive pathology. However, these graphs can only model linear interactions. Future studies could explore how these niche features evolve with therapy, potentially serving as predictive biomarkers for treatment outcomes, and should include functional investigations into the role of top-ranking genes in IBD pathogenesis.

5 Conclusion

By deconstructing colonic tissue into four cellular niches, we created a feature set capturing niche composition, spatial organization, and localized gene expression. Our model successfully addressed the three-class problem, revealing a fundamental principle of IBD pathology: subtype-specific molecular signals within these cellular niches are key to distinguishing UC from CD. In contrast, the simpler two-class task of separating IBD from HC highlighted that disruption in spatial tissue architecture is a primary indicator of general inflammation. Although we did not evaluate the model on the direct two-class problem to distinguish subtypes (UC versus CD), this represents an important next step. This study serves as a proof-of-concept, demonstrating how complex spatial data can be transformed into a robust, explainable, and predictive diagnostic framework, offering a pathway toward more precise diagnostics and targeted therapeutic strategies.

A Appendix

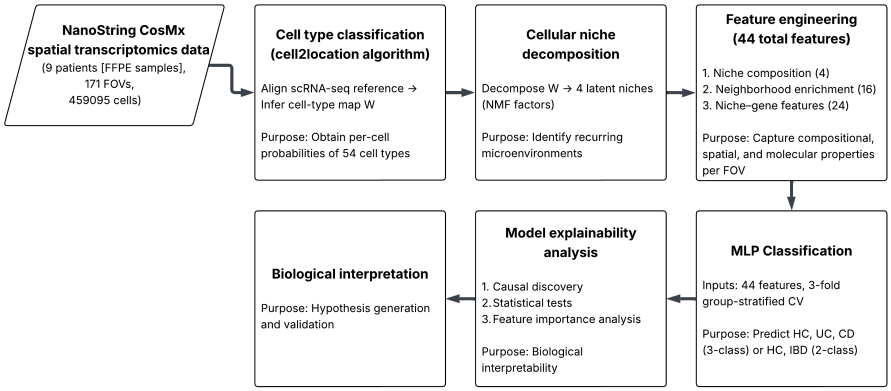
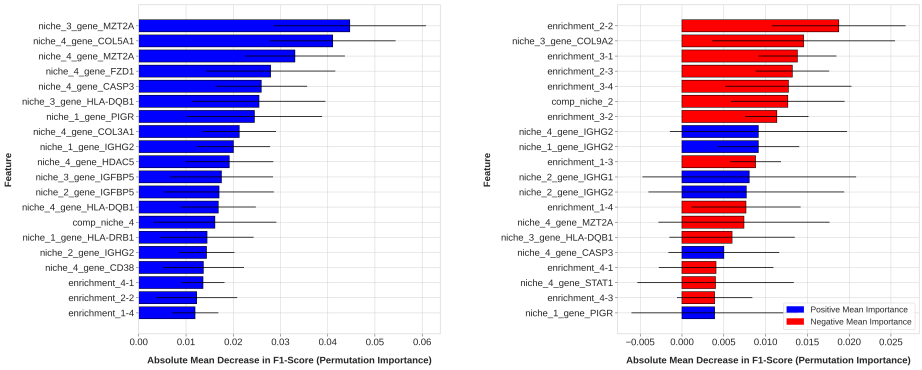


Fig. 7: Overview of the computational pipeline. Starting from NanoString CosMx spatial transcriptomics data (9 patients, 171 FOVs, 459,095 cells), cell types were inferred using the `cell2location` algorithm, and NMF to identify four recurrent cellular niches. From these niches, 44 features were engineered describing composition, spatial organization, and gene expression. A MLP classifier trained using group-stratified cross-validation was then used to predict disease states (HC vs. UC vs. CD or HC vs. IBD). Model explainability analyses (causal discovery, statistical tests, and permutation importance) enabled biological interpretation of the compositional, spatial and molecular determinants of IBD.



(a) (b)

Fig. 8: Top 20 features ranked by PI for the (a) three-class and (b) two-class problems.

Acknowledgments. This work was supported by the National Institutes of Health (NIH) grants R01HL127349, R01HL159805, R01HL178032. ChatGPT (GPT-5) was

used solely to optimize phrasing and reduce wordiness in order to meet the page limits set by the conference. All ideas, analyses, and conclusions presented in this manuscript are entirely our own.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Flynn, S., Eisenstein, S.: Inflammatory bowel disease presentation and diagnosis. Surg. Clin. N. Am. 99, 6, 1051–1062 (2019). https://doi.org/10.1016/j.suc.2019.08.001.
2. Verstockt, B., Bressler, B., Martinez-Lozano, H., et al.: Time to revisit disease classification in inflammatory bowel disease: is the current classification of inflammatory bowel disease good enough for optimal clinical management? Gastroenterology 162, 5, 1370–1382 (2022). https://doi.org/10.1053/j.gastro.2021.12.246.
3. Ungaro, R., Mehandru, S., Allen, P.B., et al.: Ulcerative colitis. Lancet 389, 10080, 1756–1770 (2017). https://doi.org/10.1016/S0140-6736(16)32126-2.
4. Kucharzik, T., Koletzko, S., Kannengiesser, K., et al.: Ulcerative colitis—diagnostic and therapeutic algorithms (17.08.2020). Dtsch. Arztebl. 117, 564–573 (2020). https://doi.org/10.3238/arztebl.2020.0564.
5. Smillie, C.S., Biton, M., Ordovas-Montanes, J., et al.: Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. Cell 178, 3, 714–730.e22 (2019). https://doi.org/10.1016/j.cell.2019.06.029.
6. Martin, J.C., Chang, C., Boschetti, G., et al.: Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. Cell 178, 6, 1493–1508.e20 (2019). https://doi.org/10.1016/j.cell.2019.08.008.
7. Elmentaite, R., Kumasaka, N., Roberts, K., et al.: Cells of the human intestinal tract mapped across space and time. Nature 597, 7875, 250–255 (2021). https://doi.org/10.1038/s41586-021-03852-1.
8. Garrido-Trigo, A., Corraliza, A.M., Veny, M., et al.: Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. Nat Commun. 14, 1, 4506 (2023). https://doi.org/10.1038/s41467-023-40156-6.
9. Gudiño, V., Bartolomé-Casado, R., Salas, A.: Single-cell omics in inflammatory bowel disease: recent insights and future clinical applications. Gut 74, 8, 1335–1345 (2025). https://doi.org/10.1136/gutjnl-2024-334165.
10. Ståhl, P.L., Salmén, F., Vickovic, S., et al.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353, 6294, 78–82 (2016). https://doi.org/10.1126/science.aaf2403.
11. Rao, A., Barkley, D., França, G.S., et al.: Exploring tissue architecture using spatial transcriptomics. Nature 596, 7871, 211–220 (2021). https://doi.org/10.1038/s41586-021-03634-9.
12. Kleshchevnikov, V., Shmatko, A., Dann, E., et al.: Cell2location maps fine-grained cell types in spatial transcriptomics. Nat. Biotechnol. 40, 5, 661–671 (2022). https://doi.org/10.1038/s41587-021-01139-4.
13. Mayr, C.H., Santacruz, D., Jarosch, S., et al.: Spatial transcriptomic characterization of pathologic niches in IPF. Sci. Adv. 10, 32, 1–15 (2024). https://doi.org/10.1126/sciadv.adl5473.

14. Parigi, S.M., Larsson, L., Das, S., et al.: The spatial transcriptomic landscape of the healing mouse intestine following damage. Nat. Commun. 13, 1, 828 (2022). https://doi.org/10.1038/s41467-022-28497-0.
15. Kong, L., Subramanian, S., Segerstolpe, Å., et al.: Single-cell and spatial transcriptomics of stricturing Crohn's disease highlights a fibrosis-associated network. Nat. Genet. 57, 7, 1742–1753 (2025). https://doi.org/10.1038/s41588-025-02225-y.
16. Mennillo, E., Kim, Y.J., Lee, G., et al.: Single-cell and spatial multi-omics highlight effects of anti-integrin therapy across cellular compartments in ulcerative colitis. Nat. Commun. 15, 1, 1493 (2024). https://doi.org/10.1038/s41467-024-45665-6.
17. Garrido-Trigo, A., Heyn, H., Salas, A.: IBD CosMx NanoString data from colonic mucosa. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE234713 (2023).
18. CosMx™ Spatial Molecular Imager, https://nanostring.com/products/cosmx-spatial-molecular-imager/, last accessed 2025/9/7
19. Garrido-Trigo, A., Salas, A.: IBD single cell data from colonic mucosa. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE214695 (2023).
20. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Proceedings of the 14th International Conference on Neural Information Processing Systems (NIPS 2000), pp. 535–541. MIT Press, Cambridge, MA (2000).
21. Boutsidis, C., Gallopoulos, E.: SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recognit. 41, 4, 1350–1362 (2008). https://doi.org/10.1016/j.patcog.2007.09.010.
22. Maneewongvatana, S., Mount, D.M.: Analysis of approximate nearest neighbor searching with clustered point sets. arXiv cs/9901013 (1999). https://doi.org/10.48550/arXiv.cs/9901013.
23. Fano, R.M.: Transmission of Information: A Statistical Theory of Communication. MIT Press, Cambridge, MA, USA (1961).
24. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18, 1, 50–60 (1947). https://doi.org/10.1214/aoms/1177730491.
25. Fay, M.P., Proschan, M.A.: Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat. Surv. 4, 1–39 (2010). https://doi.org/10.1214/09-SS051.
26. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Methodol. 57, 1, 289–300 (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
27. Almeida, L.B.: Multilayer perceptrons. In: Handbook of Neural Computation. CRC Press (1996).
28. Goodfellow, I., Bengio, Y., Courville, A.: Deep feedforward networks. In: Deep Learning, pp. 164–223. MIT Press, Cambridge, MA (2016).
29. Feurer, M., Hutter, F.: Hyperparameter optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds.) Automated Machine Learning. The Springer Series on Challenges in Machine Learning, pp. 3–33. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05318-5_1.
30. Goodfellow, I., Bengio, Y., Courville, A.: Rectified linear units and their generalizations (Subsection 6.3.1). In: Deep Learning, pp. 189–191. MIT Press, Cambridge, MA (2016).

31. Goodfellow, I., Bengio, Y., Courville, A.: Logistic sigmoid and hyperbolic tangent (Subsection 6.3.2). In: Deep Learning, pp. 191–192. MIT Press, Cambridge, MA (2016).
32. Goodfellow, I., Bengio, Y., Courville, A.: Regularization for deep learning. In: Deep Learning, pp. 224–270. MIT Press, Cambridge, MA (2016).
33. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2017). https://doi.org/10.48550/arXiv.1412.6980.
34. Opitz, J.: A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Trans. Assoc. Comput. Linguist. 12, 820–836 (2024). https://doi.org/10.1162/tacl_a_00675.
35. Spirtes, P.: An anytime algorithm for causal inference. In: Richardson, T.S., Jaakkola, T.S. (eds.) Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, pp. 278–285. PMLR (2001).
36. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. J. Mach. Learn. Res. 15, 3921–3962 (2014).
37. Lovelace, T.: rCausalMGM. GitHub repository. https://github.com/tyler-lovelace1/rCausalMGM (2025).
38. Ge, X., Raghu, V.K., Chrysanthis, P.K., et al.: CausalMGM: an interactive web-based causal discovery tool. Nucleic Acids Res. 48, W1, W597–W602 (2020). https://doi.org/10.1093/nar/gkaa350.
39. Ramsey, J.: Improving accuracy and scalability of the PC algorithm by maximizing p-value. arXiv preprint arXiv:1610.00378 (2016). https://doi.org/10.48550/arXiv.1610.00378.
40. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47, 583–621 (1952). https://doi.org/10.1080/01621459.1952.10483441.
41. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. 20, 177, 1–81 (2019).
42. Altmann, A., Toloşi, L., Sander, O., et al.: Permutation importance: a corrected feature importance measure. Bioinformatics 26, 10, 1340–1347 (2010). https://doi.org/10.1093/bioinformatics/btq134.
43. Goyette, P. et al.: High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nat Genet. 47, 2, 172–179 (2015). https://doi.org/10.1038/ng.3176.
44. Rüthlein, J. et al.: Immunoglobulin G (IgG), IgG1, and IgG2 determinations from endoscopic biopsy specimens in control, Crohn's disease, and ulcerative colitis subjects. Gut. 33, 4, 507–512 (1992). https://doi.org/10.1136/gut.33.4.507.
45. Bruno, M.E.C. et al.: Correlation of Biomarker Expression in Colonic Mucosa with Disease Phenotype in Crohn's Disease and Ulcerative Colitis. Dig Dis Sci. 60, 10, 2976–2984 (2015). https://doi.org/10.1007/s10620-015-3700-2.
46. Sarvestani, S.K. et al.: Induced organoids derived from patients with ulcerative colitis recapitulate colitic reactivity. Nat Commun. 12, 1, 262 (2021). https://doi.org/10.1038/s41467-020-20351-5.
47. Silverberg, M.S. et al.: A Population- and Family-Based Study of Canadian Families Reveals Association of HLA DRB1*0103 With Colonic Involvement in Inflammatory Bowel Disease. Inflamm Bowel Dis. 9, 1, 1–9 (2003). https://doi.org/10.1097/00054725-200301000-00001.

48. Magalhães, H.I.R. et al.: Study of the roles of caspase-3 and nuclear factor kappa B in myenteric neurons in a P2X7 receptor knockout mouse model of ulcerative colitis. World Journal of Gastroenterology. 29, 22, 3440–3468 (2023). https://doi.org/10.3748/wjg.v29.i22.3440.
49. Ning, L. et al.: Quantitative Proteomic Analysis Reveals the Deregulation of Nicotinamide Adenine Dinucleotide Metabolism and CD38 in Inflammatory Bowel Disease. BioMed Research International. 2019, 1, 3950628 (2019). https://doi.org/10.1155/2019/3950628.
50. Mizoguchi, E.: Chitinase 3–Like-1 Exacerbates Intestinal Inflammation by Enhancing Bacterial Adhesion and Invasion in Colonic Epithelial Cells. Gastroenterology. 130, 2, 398–411 (2006). https://doi.org/10.1053/j.gastro.2005.12.007.
51. Ballengee, C.R. et al.: Association Between Plasma Level of Collagen Type III Alpha 1 Chain and Development of Strictures in Pediatric Patients With Crohn's Disease. Clinical Gastroenterology and Hepatology. 17, 9, 1799–1806 (2019). https://doi.org/10.1016/j.cgh.2018.09.008.
52. Valencia, C.W. et al.: A Model to Study Intestinal Fibrosis in iPSC-Derived Human Epithelial AND Mesenchymal Cells from Crohn's Disease Patients. Inflamm Bowel Dis. 30, Supplement_1, S57 (2024). https://doi.org/10.1093/ibd/izae020.116.
53. You, J. et al.: Wnt Pathway-Related Gene Expression in Inflammatory Bowel Disease. Dig Dis Sci. 53, 4, 1013–1019 (2008). https://doi.org/10.1007/s10620-007-9973-3.
54. Chervy, M. et al.: Epigenetic master regulators HDAC1 and HDAC5 control pathobiont Enterobacteria colonization in ileal mucosa of Crohn's disease patients. Gut Microbes. 14, 1, 2127444 (2022). https://doi.org/10.1080/19490976.2022.2127444.
55. Xiong, Z. et al.: Identification and Validation of Signature Genes and Potential Therapy Targets of Inflammatory Bowel Disease and Periodontitis. JIR. 16, 4317–4330 (2023). https://doi.org/10.2147/JIR.S426004.
56. Kinchen, J. et al.: Structural Remodeling of the Human Colonic Mesenchyme in Inflammatory Bowel Disease. Cell. 175, 2, 372-386.e17 (2018). https://doi.org/10.1016/j.cell.2018.08.067.
57. Schreiber, S. et al.: Activation of signal transducer and activator of transcription (STAT) 1 in human chronic inflammatory bowel disease. Gut. 51, 3, 379–385 (2002). https://doi.org/10.1136/gut.51.3.379.
58. Rao, R.B. et al.: On the Dangers of Cross-Validation. An Experimental Evaluation. In: Proceedings of the 2008 SIAM International Conference on Data Mining (SDM). pp. 588–596 Society for Industrial and Applied Mathematics (2008). https://doi.org/10.1137/1.9781611972788.54.